Extending the DSE: LOD Support and TEI/IIIF Integration in EVT

Paolo Monella
Venice Centre for Digital and Public Humanities, Università Ca' Foscari di Venezia
Università degli Studi di Palermo
paolo.monella@unipa.it[1]

Roberto Rosselli Del Turco
Dipartimento di Studi Umanistici
Università di Torino
LabCD, Università di Pisa
roberto.rosselidelturco@unito.it

Abstract

English. Current digital scholarly editions (DSEs) have the opportunity of evolving to dynamic objects interacting with other Internet-based resources thanks to open frameworks such as IIIF and LOD. This paper showcases and discusses two new functionalities of EVT (Edition Visualization Technology), version 2: one improving the management of named entities (f.i. personal names) through the use of LOD resources such as FOAF and DBpedia; the other, providing integration of the published text with digital images of the textual primary sources accessed from online repositories (e.g. e-codices or digital libraries such as the Vaticana or the Ambrosiana) via the IIIF protocol.

Italiano. Le edizioni critiche digitali oggi hanno l'opportunità di diventare sistemi dinamici che interagiscono con altre risorse su Internet grazie a *framework* aperti come IIIF e LOD. Questo saggio mostra e discute due nuove funzionalità della versione 2 di EVT (Edition Visualization Technology): il primo migliora la gestione delle *named entities* (ad es. i nomi di persona) attraverso l'uso di risorse LOD come FOAF e DBpedia; il secondo integra il testo pubblicato con le immagini digitali delle fonti testuali, recuperate da server online (ad es. e-codices o le biblioteche digitali Vaticana e Ambrosiana) tramite il protocollo IIIF.

1 Introduction

In the currently available DSEs there is a considerable lack of homogeneity since, besides a small number of (necessarily) common features, some are still rather traditional in the way of modelling the data, of conceiving the visualization of the edited text and in the kind of tools which they make available to the scholar, while others explore the enrichment of the edition's contents and the design of research tools in a highly innovative way. As W. Gibson (1990) said "The future is already here – it's just not very evenly distributed".

As a matter of fact, the first, pioneering DSEs were quite conservative in their approach to User Interface (UI) layout. This is understandable, since the first electronic editions were considered an equivalent of traditional editions in a different medium: the "printed page paradigm" (Sahle, 2016) inspired a mimetic design which would take into account very few advantages of the new publication framework besides its ubiquity and the apparently endless space it grants. The "remediation" (Bolter & Grusin, 2000) of the scholarly edition in the new media had not yet taken place.

When the methodology advances possible thanks to a Web-based digital edition started to be evident, a new research field was born: digital philology can be traced back to the desire to explore the potential of a truly *Digital* Scholarly Edition. This impulse has led to a lively but somewhat chaotic research activity, with an apparent paradox:

- only projects which can rely on generous resources may afford to explore new approaches to a DSE design, but usually they are focused on the specific task at hand, while the decision makers aren't interested in broader reflections on the general methodology;

[1] For the purposes of the Italian academy, R. Rosselli Del Turco is responsible for sections 1-4 and P. Monella for sections 5 and 6. The initial abstract was jointly written by the two authors, who also planned and revised the article together.

- interested scholars, on the other hand, may be hampered not only by the lack of resources to experiment, but also by the fact that the DH-related IT world is moving very fast, so fast that sometimes new technologies are introduced, enhanced, exploited and set aside within a few years.

This article aims at presenting some of the latest methods which can be applied to a DSE with the purpose of making it an even better (more flexible, modular, distributed, interconnected) research tool, and will also consider the software design and implementation issues that they imply.

2 No DSE is an island

Of the many limits artificially imposed to the DSE by the printed page paradigm, the first to fall was that of the book as a monolithic product, isolated from other books and unalterable, if not with rather high costs, after its publication. A DSE, in fact, can be changed both occasionally, f.i. to correct errors, and systematically, to add new texts, commentaries, bibliographic items, etc., so that a publication date by no means implies the end of the editing process, rather just the beginning of a new phase. Furthermore, a DSE is a dynamic, not a static object, a research tool which assists the scholar in data interpretation and analysis; and it is not a "closed box", but it can engage in dialogue and interaction with other Internet-based resources thanks to the global linking framework upon which the Web itself is built (Bodard & Garcés, 2009; van Zundert & Boot, 2011).

As a consequence, the greatest advantage of Web-based publication is not only that it makes scholarly content dissemination much easier and cheaper, but that the DSE can access other resources (and be accessed), and it can rely on external assets and services for specific functionalities. Examples include:

- taking advantage of semantic web technologies and Linked Open Data to enrich the edition content;
- text/image linking, pointing to digital collections of images of manuscripts maintained by external repositories (IIIF framework);
- modelling intertextual relationships through canonical text services (CTS and DTS protocols).

While this deep interconnecting and sharing of resources may look like a "quantitative" only advantage when compared to traditional scholarly editions (i.e., you can modify your edition when you want and integrate all the available content that you may see fit), there are important methodological consequences:

- the definitive rejection of the concept of an edition as an isolated and immutable entity;
- the acknowledgement of the fact that the use of external materials to integrate a DSE is highly desirable as it results in an enrichment of the DSE itself;
- an impulse to the collaboration between different edition projects, possibly adopting principles typical of the social edition (Siemens *et al*. 2012);
- as a consequence, an incentive to adopt a modular and distributed approach in the design of digital editions, making them more flexible;
- the possibility of virtually re-assemble dismembered manuscripts scattered across several preserving institutions (see f.i. the Fragmentarium project);
- a simplification of the problem of copyright management for digital reproductions of MS images by libraries, since images are functionally integrated in the DSE, but remain resident in their repository;
- the DSE itself may become a resource for other editions if the data on which it is based is published in such a way as to make it available to third parties.

This approach, however, poses a number of problems:

- greater technical complexity: so far only the editions defined as "haute couture" by E. Pierazzo (2019) can afford to adopt this approach because its digital implementation is certainly more complex than a simple reproduction of static images and texts in a Web-based edition;
- qualitative homogeneity of the different components of a DSE: if part of the content is entrusted to external resources, it is of critical importance that these resources meet the same academic standards as the original materials;
- long term sustainability: the interdependence between the "internal" and "external" components of a

DSE makes it possible to modify the existing connections and to add new materials as soon as they become available, but it also implies a strict and continued control on the actual availability and compatibility of these materials in the long run.

What is needed on the methodological level:

- open protocols, for data-sharing infrastructures, and open licenses, for resources to be shared;
- an open and ongoing scholarly discussion on the systematization of fundamental concepts to create a shared conceptual framework: we need shared terminologies and open ontologies as a necessary methodological condition for the creation of such shared resources;
- an integration of those methodological experiments in common editorial practice.

3 EVT 2: the linked DSE

EVT 2 is the second version, currently based on the AngularJS framework, of EVT (Edition Visualization Technology), a software tool for publishing editions based on the TEI/XML format, which has been developed in such a way as to go beyond the printed page paradigm – especially with regard to the User Interface design (see Di Pietro & Rosselli Del Turco 2018). The way in which it manages data, however, makes it mostly suitable for self-contained editions, created on the basis of local resources. Furthermore, at the present moment EVT is based on the client-only architecture, which presents many advantages (it is easy to install, little to no maintenance is required, it has indefinite durability), but also quite a few limits concerning important functionality (such as server functionality, f.i. for textual searches or to serve images).

For these reasons, it is an important goal of the development team to add support for protocols such as LOD for semantic Web resources, IIIF for images and CTS/DTS for intertextual relationships, so that scholars can count on a *prêt-à-porter* tool for their work. In the long term this goal may be achieved by adding RESTFul services, to add server functionality without encumbering the software too much, but it is already possible for projects based on EVT 2 to include LOD and IIIF resources. This will also have the beneficial effect of allowing more widespread knowledge of these protocols.

4 EVT and LOD

Although LOD resources are extremely diversified in terms of type of content, semantic-based operation, etc., it is undeniable that this is a very promising area which has seen constant growth in recent years.

EVT is already able to access resources on the Web thanks to URIs specified in the TEI markup, in fact some of the existing EVT-based editions (e.g. the Codice Pelavicino Digitale) use resources such as external repositories to provide additional information about the named entities identified in the text. The management of named entities can be significantly improved through the use of LOD resources such as FOAF and DBpedia.

In the Codice Pelavicino Digitale, person names are already linked to the *Dizionario Biografico degli Italiani* when possible, for instance:

```
<!--Code snippet 1-->
<note>Guido da Velate, arcivescovo di Milano dal 1045 al 1068, morto nel 1071 (si
veda <ref target="http://www.treccani.it/enciclopedia/guido-da-velate_
%28Dizionario-Biografico%29" type="biblio">Guido da Velate nel Dizionario
Biografico degli Italiani</ref>).</note>
```

This information can be supplemented or replaced by linking to a DBpedia entry:

```
<!--Code snippet 2-->
<ref target="http://dbpedia.org/page/Guido_da_Velate">Guido da Velate</ref>
```

so that it is possible to take advantage of the wealth of connections and of the sophisticated ontologies that LOD resources allow. Furthermore, as hinted above the DSE itself can make (part of) its material available as LOD, so that other editions can build upon it. The concept of "distributed edition", therefore, is coming closer to reality, in fact this is the goal of a new research project aiming at disseminating a DSE on sustainable and public resources such as Zenodo and GitHub (see O'Donnell *et al.*, 2018; note that EVT already runs on GitHub).

5 EVT and IIIF

5.1 The rise of IIIF

One notable example of an open protocol for online resource integration is IIIF – International Image Interoperability Framework. IIIF is rapidly emerging as a technology to exchange and integrate image-based resources in Web-based systems. One interesting use-case is a DSE in which portions of a TEI-encoded text based on a primary source such as a manuscript or a printed book are linked to the images of that source, stored and described via the IIIF framework. For instance, the transcription of a page of a manuscript can be linked to its facsimile, and the transcription of a line can be linked to the region of that facsimile corresponding to the line.

5.2 IIIF Image and Presentation APIs

The IIIF protocol defines different APIs, two of which will be briefly discussed here:

1. The *IIIF Image API* "specifies a web service that returns an image in response to a standard HTTP or HTTPS request". Simply put, this API returns *one* image or a portion of it. The description on the image are stored in a JSON file named `info.json`.
2. The *IIIF Presentation API* instead "describes how the structure and layout of a complex image-based object can be made available in a standard manner". Such a complex object can be a *collection* of digital images of a manuscript, accompanied by the relevant metadata, stored in a JSON file named `manifest.json`.

5.3 TEI and IIIF: a marriage made in heaven?

The TEI approach has always been text-centric, and only more recently the TEI editors have included a document-based approach in which the digital images of a textual source have equal dignity as its textual representation, via the <facsimile> / <surface> / <zone> encoding approach. On the other hand, IIIF is overtly and intentionally image-based. The TEI/IIIF integration thus looks very promising and productive for DSEs aiming to combine textual representation and digital images. However, at this point this is very much an open field of experimentation.

5.4 Two directions for a TEI/IIIF integration

Two approaches are theoretically possible for this integration:

1. Linking from IIIF to TEI

- *How*: according to the IIIF Presentation API, the node of a IIIF manifest identifying a specific (portion of an) image can point to an *annotationList* (a separate JSON file) including an annotation pointing to an external TEI XML file with the relevant textual representation (transcription).
- *Why this might not be a good idea*: the institution curating the IIIF collection (for textual sources, most probably a library or an archive) should create, curate and update the annotations linking to the TEI XML files. If those files belong to external DSEs incorporating the IIIF images, the DSE URIs might change and require constant update from the library's side – which is clearly not sustainable. On the other side, the library could create those TEI XML files itself to store and expose transcriptions of its own manuscripts, but in the current division of labour, libraries focus on digital imaging and metadata rather than on full transcriptions and textual criticism.

2. Linking from TEI to IIIF

- *How*: within a TEI XML file, f.i. in a <pb/> (page beginning) or <lb/> (line beginning) element, a @facs attribute points to a IIIF URI, either directly or indirectly.
    - *2A - Directly*: @facs takes the relevant IIIF URI as value (identifying, f.i., a whole manuscript page or a rectangle of that page including a line);
    - *2B - Indirectly*: @facs points to a <surface> or <zone> element within the <facsimile> section of the TEI file, and the <surface> / <zone> element points to the relevant IIIF URI.

The indirect strategy adds a layer of complexity, but also increases flexibility.

- *Why this might be a good idea*: this approach keeps resources separated (TEI XML files for text-centered digital philology and a IIIF infrastructure for digital image collections), with modularity and interoperability in mind. Many DSEs can link to the same IIIF image collections. Shortly said, philologists work on text with TEI, librarians work on document digitization with IIIF.

5.5 IIIF implementation in EVT

EVT 2 now features IIIF integration thanks to OpenSeadragon, its embedded image viewer. It implements the second approach described in the previous paragraph ("Linking from TEI to IIIF") and uses the IIIF Image API. Digital philologists can thus integrate external images of a textual source, hosted by a third-party IIIF image server, in the DSE, with an arbitrary level of alignment granularity. IIIF-compliant servers include e-codices, the Veneranda Biblioteca Ambrosiana in Milan (Cusimano, 2019) or the Biblioteca Apostolica Vaticana in Rome.

More precisely, EVT currently implements encoding strategy 2A described in paragraph 5.4 above:

1. The TEI XML source code has an element pointing to an image exposed by a IIIF server (typically a facsimile of a page) or to a portion of that image (typically a rectangle including a line or an other textual division). In the following code sample, we are pointing to a *whole image* (representing a manuscript page) from the IIIF server *e-codices* - *Virtual Manuscript Library of Switzerland*. The value of @facs is a URI following the IIIF Image API:

```
<!--Code snippet 3-->
<pb facs="https://www.e-codices.unifr.ch/loris/csg/csg-0730/csg-0730_002.jp2/
         full/full/0/default/jpg"/>
```

2. The image viewer integrated in EVT, OpenSeadragon, dereferences the URI from the @facs attribute, fetches the image from the external IIIF server and shows the whole image of the manuscript page alongside its TEI-based transcription.

Please note that the version of OpenSeadragon currently embedded in EVT (2.4.1) also allows to align a <pb/> element with a *specific portion of an image*, defined as a rectangle as per the IIIF Image API. Thus code snippet 3 above can be edited to:

```
<!--Code snippet 4-->
<pb facs="https://www.e-codices.unifr.ch/loris/csg/csg-0730/csg-0730_002.jp2/
         1800,600,3000,5000/full/0/default/jpg"/>
```

to crop out the manuscript page margins. Coordinates "1800,600" (in pixels) define the top left corner of the rectangle, "3000" defines the rectangle's base, "5000" its height. The encoding strategies described so far fulfill the common need of editors to pair <pb/> elements with a manuscript or book page (as well with the surface of an inscription, a tablet or any other support) and to display it aside the transcription.

The following code sample, instead, exemplifies the TEI XML encoding strategy currently supported by EVT to link elements such as <lb/>, <p> or to smaller portions of the surface image:

```
<!--Code snippet 5-->
<surface>
  <zone lrx="1052" lry="211" rend="visible" rendition="Line" ulx="261" uly="156"
           xml:id="zone-line-2-1"/>
</surface>
[...]
  <pb facs="https://www.e-codices.unifr.ch/loris/csg/csg-0730/csg-0730_002.jp2/
           full/full/0/default/jpg"/>
  <lb facs="#zone-line-2-1"/>
```

In code snippet 5 (which incorporates snippet 3 above), the <pb/> element references the whole manuscript page via a IIIF URI; <lb/> points to a <zone> element in the <facsimile> section that defines a rectangle within the IIIF image through the internal TEI XML encoding strategy: @lrx and @lry define the coordinates of the lower right corner of the rectangle, @ulx and @uly those of the and upper left corner (see attribute class att.coordinated in the P5 TEI *Guidelines*). Please note that this is different than strategy 2B from paragraph 5.4 because it still is the <pb/> element, not <zone>, that is retrieving the IIIF image.

5.6 Future development

Support for encoding strategy 2B (<pb/> or <lb/>'s attribute @facs points to <surface> or <zone>, and the latter points to an image in an IIIF server) is not yet available in EVT, but the development team aims at including it in future releases. These are examples of TEI XML code that will be managed by EVT:

```
<!--Code snippet 6-->
<facsimile>
  <surface xml:id="image-p2">
  <graphic url="https://www.e-codices.unifr.ch/loris/csg/csg-0730/csg-
0730_002.jp2/
               full/full/0/default/jpg"/>
  <!--'zone' elements may be included here-->
  </surface>
</facsimile>
[...]
<pb facs="#image-p2"/>
```

Or, with a more compact encoding:

```
<!--Code snippet 7-->
<facsimile>
  <surface xml:id="image-p2" facs="https://www.e-codices.unifr.ch/loris/csg/
                  csg-0730/csg-0730_002.jp2/full/full/0/default/jpg">
  <!--'zone' elements may be included here-->
  </surface>
</facsimile>
[...]
<pb facs="#image-p2"/>
```

Another feature that may be implemented in the future, should any project collaborating with the EVT team express this need, is the definition of a rectangle within an image (e.g. for a manuscript line encoded with <lb/>) not through the TEI XML strategy (attributes @lrx, @lry, @ulx and @uly in <zone>), but directly through the IIIF image API, such as in the following code:

```
<!--Code snippet 8-->
<lb facs="https://www.e-codices.unifr.ch/loris/csg/csg-0730/csg-0730_002.jp2/
         1900,930,2580,140/full/0/default/jpg">
```

Finally, an experimental version of EVT based on the Angular 8 framework, not yet available for download, already supports the IIIF Presentation API and loads the full `manifest.json` document description including metadata on all manuscript images, while the image API currently supported by EVT provides access to one manuscript page image at a time. To load a whole `manifest.json` file, it will be enough to specify its URL in the EVT `config.json` configuration file, e.g.

```
<!--Code snippet 9-->
{ "title": "My Digital Edition",
  "manifestURL": "https://www.e-codices.unifr.ch/metadata/iiif/csg-
0730/manifest.json" }
```

Besides allowing a quick publication of all pages of a manuscript, this is a first step towards exploiting the full potential and flexibility of the IIIF framework.

6. Conclusions

The EVT development team is committed to supporting the current trend towards the distributed DSE, integrating resources such as entity or relationship definitions (LOD) and images (IIIF), as well as text fragments (CTS/DTS) in future versions. Internal and external objects alike can be better modelled by the new modular and object-oriented implementation adopted in EVT 2. Further development aims at supporting alternative encoding strategies for linking to external resources while keeping the whole edition (XML, other internal and external data, software) sustainable, durable and compliant with the FAIR principles, according to which "all research objects should be Findable, Accessible, Interoperable and Reusable (FAIR) both for machines and for people" (Wilkinson *et al.*, 2016).

References

Gabriel Bodard and Juan Garcés. 2009. Open Source Critical Editions: A Rationale. In M. Deegan & K. Sutherland (eds.), *Text Editing, Print, and the Digital World* 83–98. Ashgate Publishing, Aldershot.

Jay D. Bolter and Richard A. Grusin. 2000. *Remediation: understanding new media*. MIT Press, Cambridge (Mass.) and London.

Fabio Cusimano. 2019. Biblioteche di conservazione & Data Curation: dal Custos catalogi al Digital Librarian. Il caso della Veneranda Biblioteca Ambrosiana. *JLIS* 10,1:125-139. DOI: http://dx.doi.org/10.4403/jlis.it-12513

Chiara Di Pietro and Roberto Rosselli Del Turco. 2018. Between Innovation and Conservation: The Narrow Path of User Interface Design for Digital Scholarly Editions. In *Digital Scholarly Editions as Interfaces*, edited by Roman Bleier, Martina Bürgemeister, Helmut W. Klug, Frederike Neuber, and Gerlinde Schneider. *Schriften Des Instituts Für Dokumentologie Und Editorik* 12: 133–63. BoD, Norderstedt. https://kups.ub.uni-koeln.de/9085/

William Gibson. 1990. *Cyberpunk* (Documentary). Directed by Marianne Trench, produced by Peter von Brandenburg, An Intercon Production. [Excerpt occurs in ~~Part 3 of~~ 5 parts; Timecode 12:20 of 14:59] (Video available in 5 parts on Youtube: https://www.youtube.com/watch?v=xxTuEGEl9EQ).

Lauren Magnuson. 2016. Store and display high resolution images with the International Image Interoperability Framework (IIIF). *ACRL TechConnect Blog*. URL: https://acrl.ala.org/techconnect/post/store-and-display-high-resolution-images-with-the-international-image-interoperability-framework-iiif/

Daniel Paul O'Donnell, Gurpreet Singh, Dot Porter, Roberto Rosselli Del Turco, Marco Callieri, Matteo Dellepiane, and Roberto Scopigno. 2018. Publishing (and Forgetting) the Small or Medium-sized Scholarly Edition or Cultural Heritage Collection as Linked Open Data: Using Zenodo and Github to Publish the Visionary Cross Project (Abstract).
URL: https://zenodo.org/record/3338482#.XeDjqdF7mV4. DOI: http://doi.org/10.5281/zenodo.3338482

Elena Pierazzo. 2019. What future for digital scholarly editions? From Haute Couture to Prêt-à-Porter. *International Journal of Digital Humanities*. https://doi.org/10.1007/s42803-019-00019-3

Roberto Rosselli Del Turco, Giancarlo Buomprisco, Chiara Di Pietro, Julia Kenny, Raffaele Masotti and Jacopo Pugliese. 2015. Edition Visualization Technology: A Simple Tool to Visualize TEI-Based Digital Editions. *Journal of the Text Encoding Initiative*. Issue 8. URL: http://jtei.revues.org/1077; DOI: https://doi.org/10.4000/jtei.1077

Roberto Rosselli Del Turco (ed.). 2017. *The Digital Vercelli Book. A facsimile edition of Vercelli, Biblioteca Capitolare, CXVII*. Transcription and encoding by Roberto Rosselli Del Turco, Raffaele Cioffi, Federica Goria. EVT software created by Chiara Di Pietro, Julia Kenny, Raffaele Masotti, Roberto Rosselli Del Turco. *Collane@unito.it*. URL: https://www.collane.unito.it/oa/items/show/11. ISBN 9788875901073.

Roberto Rosselli Del Turco. 2019. Designing an advanced software tool for Digital Scholarly Editions: The inception and development of EVT (Edition Visualization Technology). *Textual Cultures* 12.2: 91–111. URL: https://scholarworks.iu.edu/journals/index.php/textual/article/view/27690; DOI: 10.14434/textual.v12i2.27690.

Patrick Sahle. 2016. What Is a Scholarly Digital Edition? In *Digital Scholarly Editing: Theories and Practices*, edited by Elena Pierazzo and Matthew J. Driscoll. Open Book Publishers https://www.openbookpublishers.com/product/483/digital-scholarly-editing--theories-and-practices

Alberto Salarelli. 2017. International Image Interoperability Framework (IIIF): A panoramic view. *Italian Journal of Library, Archives and Information Science = Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione*, 8.

Enrica Salvatori, Edilio Riccardini, Laura Balletto, et al. (eds.). 2014. Codice Pelavicino. Edizione digitale. URL: http://pelavicino.labcd.unipi.it. ISBN 978-88-902289-0-2; DOI: https://10.13131/978-88-902289-0-2.

Ray Siemens, Meagan Timney, Cara Leitch, Corina Koolen, and Alex Garnett, with the ETCL, INKE, and PKP Research Groups. 2012. Toward modeling the social edition: An approach to understanding the electronic scholarly edition in the context of new and emerging social media. *Literary and Linguistic Computing*, 27(4): 445–461. https://doi.org/10.1093/llc/fqs013

TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. 3.6.0. Last updated on 16th July 2019. TEI Consortium. http://www.tei-c.org/Guidelines/P5/

Mark D. Wilkinson *et al.* 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3. https://www.nature.com/articles/sdata201618

Joris van Zundert, J. and Peter Boot. 2011. The digital edition 2.0 and the digital library: Services, not resources. *Digitale Edition Und Forschungsbibliothek (Bibliothek Und Wissenschaft)* 44: 141–152.